Challenges in the Processing of Audio Channels for Ambient Assisted Living

Michel Vacher, François Portet, Anthony Fleury, Norbert Noury

▶ **To cite this version:**

Michel Vacher, François Portet, Anthony Fleury, Norbert Noury. Challenges in the Processing of Audio Channels for Ambient Assisted Living. IEEE HealthCom 2010 - 12th International Conference on E-health Networking, Application & Services, Jul 2010, Lyon, France. pp.330-338. hal-00503243

HAL Id: hal-00503243
https://hal.science/hal-00503243

Submitted on 18 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Challenges in the Processing of Audio Channels for Ambient Assisted Living

Michel Vacher, François Portet
LIG Laboratory Team GETALP
UMR CNRS/UJF/INPG/UPMF 5217
385, avenue de la Bibliothèque
F-38041 Grenoble Cedex 9, France
E-mail: Michel.Vacher@imag.fr
Francois.Portet@imag.fr

Anthony Fleury
Univ. Lille Nord de France
F-59000 Lille, France
EMDouai, IA
F-59500 Douai, France
Email: Anthony.Fleury@mines-douai.fr

Norbert Noury
INL-INSA Lyon Lab. Team MMB
UMR CNRS/ECL/INSA/UCBL 5270
Av. Einstein, F-69621 Villeurbanne, France
& TIMC-IMAG Lab. Team AFIRM
Faculté de Médecine de Grenoble
Email: Norbert.Noury@insa-lyon.fr

***Abstract*—One of the greatest challenges in Ambient Assisted Living is to design *health smart homes* that could be able to anticipate the needs of its inhabitant while maintaining their comfort and their safety with an adaptation of the house environment and a facilitation of the connections to the outside world. The most likely to benefit from these smart homes are people in loss of autonomy such as the disabled people or the elderly with cognitive deficiencies. But it becomes essential to ease the interactions with the smart home through dedicated interfaces, in particular, thanks to systems reactive to vocal orders. Audio recognition is also a promising way to ensure more safety by contributing to detection of distress situations. This paper presents the stakes and the challenges of this domain based on some experiments carried out concerning distress call recognition and sound classification at home.**

I. Introduction

Evolutions in ICT led to the emergence of health smart homes designed to improve daily living conditions and independence for the population with loss of autonomy. One of the greatest challenges to be addressed by smart homes is to allow disabled and the growing number of elderly people to live independently as long as possible, before moving to a care institution. The final goal is to allow care institutions to cater for only the most severely dependent people while they are nowadays overflowed by patients due to the demographic evolutions and the rise of life expectancy. Independent living also reduces the cost to society of supporting people who have lost some autonomy.

Health smart homes started to be designed more than ten years ago and are nowadays a very active research area [1]. Three major goals are targeted. The first is to assess how a person copes with her loss of autonomy by continuous monitoring of her activities through sensor measurements. The second is to ease daily living by compensating one's disabilities (either physical or mental) through home automation. Examples include automatic light control and events reminder. The third one is to ensure security by detecting distress situations such as fall that is a prevalent fear regarding elderly persons.

To achieve these goals, smart homes are typically equipped with many sensors perceiving different aspects of the home environment. An interesting but rarely employed modality is the audio channel. Indeed, audio sensors can capture information about sounds in the home (e.g., object falling, washing machine spinning...) and about sentences that have been uttered. Speaking being the most natural way of communicating, it is thus of particular interest in distress situations (e.g., call for help) and for home automation (e.g., voice commands). More generally, vocal interfaces are much more adapted to people who have difficulties in moving than tactile interfaces (e.g., remote control) which require physical interaction. However, audio analysis in smart home is a difficult task with numerous challenges and which has rarely been deployed in real settings.

In this paper, we present the stakes and the difficulties of this task through experiments carried out in realistic settings concerning sound and speech processing for activity monitoring and distress situations recognition. The remaining of the paper is organized as follow. Related works in the audio processing for assisted living are introduced in Section II. Section III describes the AuditHIS system developed in the GETALP team for multi-source sound and speech processing. In Section IV, the results of two experiments in a smart home environment, concerning distress call recognition and activity of daily living classification, are summarized. The sounds collected in the latter constitute a precious every day life sound corpus which is described in Section V-A. Based on the analysis of this corpus and our experience, we drawn, in Section V-B, the most challenging technical issues that need to be addressed for successful development of audio processing technologies in health smart home.

II. Audio Analysis in Health Smart Homes

Audio processing, and particularly speech processing, has been a research area since the early age of Artificial Intelligence. Many methods and signal features have been explored and current state of the art techniques heavily rely on machine learning of probabilistic models (neural networks, learning vector quantization, Hidden Markov Models...). Recent developments gave impressive results and permitted speech recognition to become a feature of many industrial products, but there are many challenges that are still to be overcome

to make this modality available in health smart homes. Two major applications of audio processing in smart home have been considered: Sounds recognition to identify human-to-environment interaction (e.g., door shutting) or device functioning (e.g., washing machine), and speech recognition for vocal commands and dialogue.

Regarding sound identification, a variety of research projects applied it to assess the health status of persons living in smart homes (e.g., activity recognition or distress situations detection). For instance, sound processing can be used to quantify water usage (hygiene and drinking) [2]. Chen *et al.* [3] analyzed the audio signal via Hidden Markov Models from the Mel-Frequency Cepstral Coefficients (MFCC) to determine the different uses of the bathroom in order to recognize daily living patterns. Among theses various applications, the one that brings the most interest is the fall detection. For example, Litvak *et al.* [4] placed an accelerometer and a microphone on the floor to detect the fall of the occupant of the room by analyzing mixed sounds and vibrations. As far as assistance is concerned, the recognition of non-speech sounds associated with their direction is applied with the purpose of using these techniques in an autonomous mobile surveillance robot [5].

Regarding automatic speech recognition (ASR), some studies aimed at assisting elderly people that are not familiar with keyboards through the use of vocal commands [6]. Regarding compensation and comfort, the feasibility to control a wheel chair using a given set of vocal commands [7] was demonstrated. Moreover, microphones integrated in the ceiling of the flat for sound and speech recognition (based on Support Vector Machines with MFCC as features) aimed at achieving home automation [8].

This short overview shows the high potential of the audio channel although this modality was considered by only few of the numerous projects in the health smart home domain. The projects which studied sound detections for different aims make the research techniques scattered and difficult to standardize. Automatic speech recognition is much more focused and is essentially used for vocal command. However, their applications are rare and mostly English centered. Moreover, ASR should be adapted to the user population which is mainly composed of elderly persons. The evolution of human voice with age was extensively studied [9] and it is well known that ASR performance diminishes with growing age. Furthermore, ASR systems have reached good performances with close-talking microphone (e.g., head-set), but the performances degrade significantly as soon as the microphone is moved away from the mouth of the speaker (e.g., when the microphone is set in the ceiling). This degradation is due to a broad variety of effects including background noise and reverberation. All these problems should be taken into account in the home assisted living context.

Recently, we described AuditHIS, a complete real-time multisource audio analysis system which processes speech and sound in smart home [10]. This system has been evaluated in different realistic settings and permitted us to identify the main challenges to overcome to make audio analysis a major improvement of health smart homes.

III. The Real-time Audio Analysis System

A. The AuditHIS System

According to the results of the DESDHIS project, everyday life sounds can be automatically identified in order to detect distress situations at home [11]. Therefore, the AuditHIS software was developed to insure on-line sound and speech recognition. Figure 1-a depicts the general organization of the audio analysis system; for a detailed description, the reader is refereed to [10]. Succinctly, each microphone is connected to an input channel of the acquisition board and all channels are analyzed simultaneously. Each time the energy on a channel goes above an adaptive threshold, an audio event is detected. It is then classified as daily living sound or speech and sent either to sound classifier or to the ASR called RAPHAEL. The system is made of several modules in independent threads, synchronized by a scheduler: acquisition and first analysis, detection, discrimination, classification, and, finally, message formatting. A record of each audio event is kept and stored on the computer for further analysis. Fig. 1-b presents a screenshot of the graphical user interface.

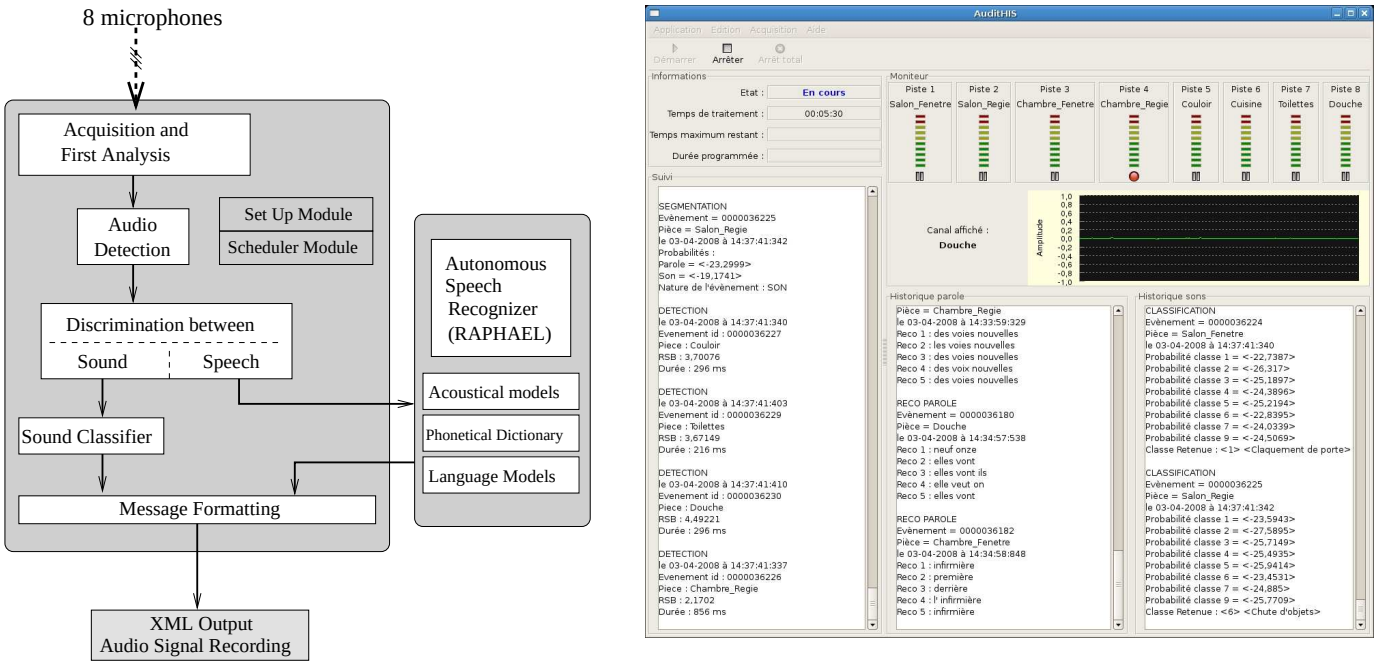
Data acquisition is operated on the 8 input channels simultaneously at a 16 kHz sampling rate. The noise level is evaluated by the first module to assess the Signal to Noise Ratio (SNR) of each acquired sound. The SNR of each audio signal is very important for the decision system to estimate the reliability of the corresponding analysis output. The detection module detects beginning and end of audio events using a adaptive threshold computed using an estimation of the background noise.

The discrimination module is based on Gaussian Mixture Model (GMM) and classifies each audio event as everyday life sound or speech. The discrimination module was trained with an everyday life sound corpus and with the Normal/Distress speech corpus recorded in our laboratory. Then, the signal is transferred by the discrimination module to the speech recognition system or to the sound classifier depending on the result of the first decision. Everyday life sounds are classified with another GMM classifier whose models were trained on an eight-class everyday life sound corpus.

B. The Autonomous Speech Analyzer

The autonomous speech recognizer RAPHAEL is running as an independent application which analyzes the speech events sent by the discrimination module. The training of the acoustic models was made with large corpora in order to ensure speaker independence. These corpora were recorded by 300 French speakers in the CLIPS (BRAF100) and LIMSI laboratories (BREF80 and BREF120). Each phoneme is then modeled by a tree state Hidden Markov Model (HMM).

Our main requirement is the correct detection of a possible distress situation through keyword detection, without understanding the person's conversation. The language model is a set of n-gram models which represents the probability of observing a sequence of n words. Our language model is



(a) General Organisation

(b) Interface of the Implemented Software

Fig. 1. The Audio Analysis System: AuditHIS

made of 299 unigrams (299 words in French), 729 bigrams (sequence of 2 words) and 862 trigrams (sequence of 3 words) which have been learned from a small corpus of French colloquial sentences (e.g., “À demain”, “J’ai bu ma tisane”...), distress phrases (e.g., “Au secours” (Help)) and home automation orders (e.g., “monte la temperature”). This small corpus is made of 415 sentences: 39 home automation orders, 93 distress sentences and 283 usual sentences.

IV. Experimentation in real conditions

The AuditHIS system has been tested in real conditions in the Health Smart Home (HIS) [12] of the TIMC-IMAG laboratory. This smart home served for the two experiments described in the remaining of this section. It is a flat of 47 m^2 inside the faculty of medicine of Grenoble. This flat is equipped with several sensors (Presence infra-red, contact door, microphones) and comprises as shown in Figure 2 all the rooms of a classical flat. Only the audio data recorded by the 8 microphones have been used in the experiments.

A. Distress Call Analysis

To assess the potential of AuditHIS to detect distress keywords in an uncontrolled environment an experiment has been run in the HIS. The aim was to test whether ASR using a small vocabulary language models was able to detect distress sentences without understanding the entire person’s conversation.

1) Experimental Set-up: Ten native French speakers were asked to utter 45 sentences (20 distress sentences, 10 normal sentences and 3 phone conversations made up of 5 sentences each). The participants included 3 women and were 37.2 (SD= 14) years old (weight: 69 $\pm$ 12 kg, height: 1.72 $\pm$ 0.08 m). The experiment took place during daytime, hence we did not control the environmental conditions of the experimental session (such as noises occurring in the hall outside the flat). The participants were situated between 1 and 10 meters away from the microphones, sat down or stood up, and have no instruction concerning their orientation with respect to the microphones (they could choose to turn their back to them). Microphones were set on the ceiling and directed vertically to the floor. The phone was placed on a table in the living room.

The participants were asked to perform a little scenario. They had to move to the living room, to close the door and then to go to the bed room and to read 30 sentences containing 10 normal and 20 distress sentences. Afterwards, they had to go to the living room and utter 30 other sentences. At the end of the scenario, each participant was called over the phone 3 times and had to read the phone conversation given (5 sentences each).

Every audio event was processed on the fly by AuditHIS and stored on the hard disk. For each event, an XML file was generated, containing the important information. During this experiment, 3164 audio events were collected with an average SNR of 12.65 dB (SD=5.6). These events do not includes 2019 ones which have been discarded because the SNR was inferior to 5 dB. This 5 dB threshold was chosen based on an empirical analysis [11].

The events were then furthermore filtered to remove duplicate instances (same event recorded on different microphones), non speech data (e.g., sounds) and saturated signal. At the end, the recorded speech corpus (7.8 minutes of signal) was

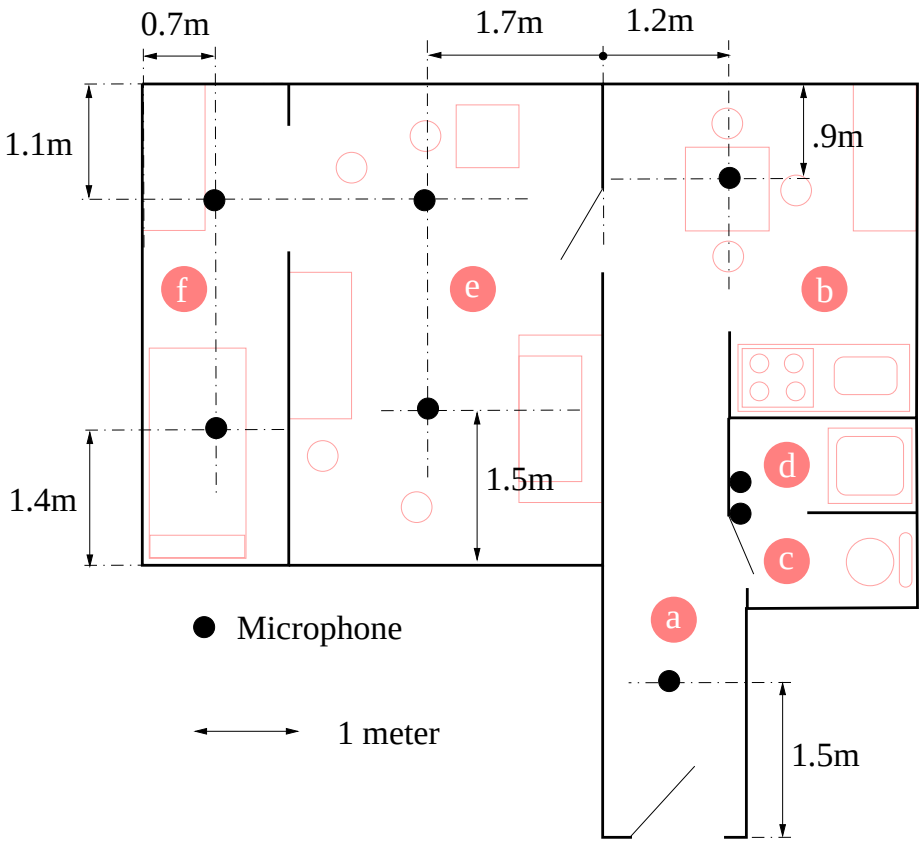


Fig. 2. Setting of the Microphones in the Health Smart Home of TIMC-IMAG, Bâtiment Jean Roget, Faculté de Médecine de Grenoble. Rooms are: (a) a corridor, (b) a kitchen, (c) toilets, (d) a bathroom, (e) a living room and (f) a bedroom.

composed of 197 distress keyword sentences (nDS) and 232 normal sentences (nNS). This corpus was indexed manually because each speaker did not follow strictly the instructions given at the beginning of the experiment.

2) Results: The 429 sentences were processed by the RAPHAEL speech recognizer using the acoustic models and the language model presented in Section III-B. To measure the performances, the Missed Alarm Rate (MAR), the False Alarm Rate (FAR) and the Global Error Rate (GER) were defined as follow:

$$MAR = \frac{nMA}{nDS}\ ,\ FAR = \frac{nFA}{nNS}\ ,\ GER = \frac{nMA + nFA}{nDS + nNS} \quad (1)$$

The results are shown in Table I. The FAR is low whatever the person. The MAR and GER, from about 5% to above 50%, highly depend on the speaker. The worst performances were observed with a speaker who uttered distress sentences like an actor. This utterance provoked variation in intensity which provoked the French pronoun “je” to not be recognized at the beginning of some sentences. For another speaker, a woman, the MAR is upper 40%. It can be explained by the fact that she walked when she uttered the sentences and made noise with her high-heeled shoes. This noise was added to the speech signals that were analyzed. The distress sentence “help” was well recognized when it was uttered with a French pronunciation but not with an English one because the phoneme [h] does not exist in French. When a sentence was uttered in the presence of an environmental noise or after a tongue clicking, the first phoneme of the recognized sentence was preferentially a fricative or an occlusive and the recognition process was altered.

TABLE I
DISTRESS KEYWORD PERFORMANCE

Speaker	MAR (%)	FAR (%)	GER (%)
1	19.0	0.0	8.9
2	5.3	0.0	2.3
3	35.0	8.7	21.0
4	16.7	4.2	9.5
5	55.0	4.5	28.6
6	36.8	4.2	18.6
7	23.5	4.3	12.5
8	42.9	5.0	24.4
9	33.3	0.0	15.5
10	24.0	8.7	16.0
Overall	29.5	4.0	15.6

The experiment led to mixed results. For half of the speakers, the distress sentences classification was correct ($FAR < 5\%$) but the other cases showed less encouraging results. This experience showed the dependence of the ASR to the speaker (thus a need for adaptation), but most of the problems were

due to noise and environmental perturbations. To investigate what problem could be encountered in health smart homes another study has been run in real setting focusing on lower aspect of audio processing rather than distress/normal situation recognition.

B. Audio Processing of Daily Living Sounds and Speech

To test the AuditHIS system, an experiment was run to acquire data in the Health Smart Home. To ensure that the data acquired would be as realistic as possible the participants were asked to perform usual daily activities. Seven activities, from the index of independence in Activities of Daily Living (ADL) were performed at least once by each participant in the HIS. These activities include: (1) Sleeping; (2) Resting: watching TV, listening to the radio, reading a magazine…; (3) Dressing and undressing; (4) Feeding: realizing and having a meal; (5) Eliminating: going to the toilets; (6) Hygiene activity: washing hands, teeth …; and (7) Communicating: using the phone. Therefore, this experiment allowed us to process realistic and representative audio events in conditions which are directly linked to usual daily living activities. The ADL scale is used by geriatricians for autonomy assessment which questions the person's ability to realize different tasks (doing a meal, hygiene, going to the toilet …). It is thus of high interest to make audio processing performing to monitor these tasks and to contribute to the assessment of the person's degree of autonomy.

1) Experimental Set up: Fifteen healthy participants (including 6 women) were asked to perform these 7 activities without condition on the time spent. Four participants were not native French speakers. The average age was 32 ± 9 years (24-43, min-max) and the experiment lasted in minimum 23 minutes 11s and 1h 35minutes 44s maximum. A visit, before the experiment, was organized to ensure that the participants will find all the items necessary to perform the seven ADLs. Participants were free to choose the order with which they wanted to perform the ADLs to avoid repetitive patterns. For more details about the experiment, the reader is refereed to [12].

It is important to note that this flat represents an hostile environment for information acquisition similar to the one that can be encountered in real homes. This is particularly true for the audio information. The sound and speech recognition system presented in [10] was tested in laboratory with an average Signal to Noise Ratio (SNR) of 27dB. In the smart home, the SNR felt to 11dB. Moreover, we had no control on the sound sources outside the flat, and there was a lot of reverberation inside the flat because of the 2 important glazed areas opposite to each other in the living room.

2) Results: The sound/speech discrimination is important for two reasons: 1) these two kinds of signal might be analyzed by different paths in the software; and 2) the fact that an audio event is identified as sound or as speech indicates very different information on the person's state of health or activity. The results of the sound/speech discrimination stage of AuditHIS are given on Table II. 2555 sounds has been detected and processed. The table shows the high confusion between the two classes. This has led to poor performance in each of the classifier (ASR and sound classification).

TABLE II
SOUND/SPEECH CONFUSION MATRIX

Target / Hits	Everyday Sound	Speech
Everyday Sound	1745 (87%)	252 (23%)
Speech	141 (25%)	417 (75%)

Some sounds like Dishes are very often confused with Speech because the training set does not include enough examples. The presence of a fundamental frequency in the spectral band of speech explains the error of the discrimination module. Falls of objects and Speech were often confused with Scream. The misclassification can be related to a design choice. Indeed, Scream is both a sound and a speech and difference between these two categories is sometimes thin. For example "Aïe!" is an intelligible scream that has been learned by the ASR but a scream could also consist in "Aaaa!" which, in this case, should be handled by the sound recognizer.

However, most of the poor performances can be explained by the too small training set and the fact that unknown and unexpected classes (e.g., thunder, Velcro) were not properly handled by the system. Our next goal is to extend this set to include more examples with more variation as well as more classes (such as water sounds and a reject class for unknown or non frequent sounds). This is mandatory because a probabilistic approach is used and a high number of instances is requested to learn correctly each class.

These results are quite disappointing but the data collected during the experiment represents a precious corpus for a better understanding of the challenges to audio processing in smart home as well as for empirical test of the audio processing models in real settings. These points are detailed in the next section.

V. AUDIO PROCESSING OF DAILY LIVING SOUNDS AND SPEECH: A CHALLENGING APPLICATION

As shown by the previous results, the processing of audio signals is a very challenging area. Many issues going from the treatment of noise and source separation to the adaptation of the model to the user and its environment need to be dealt with. Though fairly disappointing, the conducted experiments permitted to acquire a precious corpus in real conditions which has been carefully annotated. The most salient aspects of this corpus are described below in the Section V-A. Based on the preliminary analysis of this corpus and on the literature, we drawn, in Section V-B, the main challenges that audio processing in smart home needs to address before being integrable and useful to health monitoring and assistance.

A. Details of the Corpus of sounds of daily living

During the daily living experiment (Section IV-B), 1886 individual sounds and 669 sentences were collected. These were manually annotated. The most important characteristics of this corpus are summarized in Table III.

The mean SNR of each sound class is between 5 and 15 dB, far less than the SNR obtained in laboratory environment. This confirms that audio data acquired in the the health smart home were noisy as reported in Section IV-B.

The sounds acquired were very diverse much more than what we expected in an experimental condition were participants, though free to perform activities as they wanted, had a few number of recommendations to follow.

The sounds have been gathered into classes of daily living sounds according to their origin and nature. The first class is constituted of all sounds that the human body can generate. Speech apart, most of them are of low interest for the moment. However, whistling and song can be related to the mood of the person while cough and throat roughing may be associated to an health problem.

The most populated class of sound is the one related to the object and furniture handling in the house. The distribution is highly unbalanced and it is unclear how these sounds can be related to health status or distress situation. However, as shown in [13] they contribute to the recognition of activities of daily living which are essential to monitor the person's activity. Related to this class, though different, were sounds provoked by devices, such as the phone.

The most surprising class was the sound coming from the exterior of the flat (helicopter, rain, elevator, noise in the corridor...). This flat has poor noise insulation (as many homes) and we did not prevent participants any action. Thus, some of them opened the window during the experiment which was particularly annoying considering that the helicopter spot of the local hospital is at short distance. Furthermore, one of the recordings was realized during rain and thunder which artificially increased the number of sounds. Also, noise from the building disturbed the audio system because the bathroom is just near the elevator shaft.

A large number of mixed sounds also composes this corpus. Indeed, it is common that a person generates several kinds of sounds during one action. One of the most frequent case is the mixing of foot step, door closing and locker. This is probably due to the fact that participants were young and were moving quickly. This case may be less frequent with aged persons. Unclassifiable sounds were also numerous and mainly due to situations in which video were not enough to mark up with hundred percent certainty the noise occurring on the audio channel. Indeed even for a human listener, context in which a sound occurs is often essential to recognize it.

It is very important to notice that, despite the duration of the experience, the number of recorded sounds is low and highly unbalanced for the majority of the classes. Thus, the record of a sufficient number of sounds needed for statistical analysis method will be a hard task. Moreover, to acquire more generic models, it will be necessary to collect sounds in several different environments. Another important characteristic is that it is hard to annotate sounds with high certainty. It is also difficult to know which level of detail is required. The corpus contains many sounds that can be seen as super class of others (Objects shocking, Exterior ...). Moreover it is very difficult to recognize the source of the sound, but it may be of great interest to classify the sounds according to their own characteristics: periodicity, fundamental frequency, impulsive or wide spectrum... Thus, classification methods such as hierarchical or structured classification may be more adapted than flat concept modeling [14]. Finally, it is striking to see that 15% of the sounds are not directly classifiable (mixed and unknown sounds). Classification methods should thus consider ways of taking ambiguity into account and reject class.

B. Challenges that need to be addressed

Though imperfect and noisy, the data set presented in Section V-A is sufficiently large and realistic to serve an empirical analysis of the main challenges audio analysis systems must address to make smart home useful for the aging population.

1) Audio Acquisition and Processing in the Smart Home Context: In real home environment the audio signal is often perturbed by various noise (e.g., music, work on the street...). Three main sources of errors can be considered:

1) The measurement errors which are due to the position of the microphone(s) with regard to the position of the speaker;
2) The acoustic of the flat;
3) The presence of undetermined background noise such as TV or devices.

In our experiment, 8 microphones were set in the ceiling. This led to a global cover of the area but prevented from an optimal recording of speech because the individuals tend to speak horizontally. Moreover, when the person was moving, the intensity of the speech or sound changed and influenced the discrimination of the audio signals between sound and speech; the intensity change provoked as well saturation of the signal (door slamming, person coughing close to the micro). One solution could be to use head set, but this would be a too intrusive change of way of living for aging people. Though annoying, these problems are mainly perturbing for fine grain audio analysis but can be bearable in many settings.

The acoustic of the flat is another difficult problem to cope with. In our experiment, the double glazed area provoked a lot of reverberation. Similarly, every sound recorded in the toilet and bathroom area was echoed. These examples show that a static and dynamic component of the flat acoustic must be considered. Finding a generic model to deal with these issues adaptable to every home is a very difficult challenge and we are not aware of any existing solution for smart home. Of course, in the future, smart homes could be designed specifically to limit these effects but the current smart home development cannot be successful if we are not able to handle these issues when equipping old-fashioned or poorly insulated home.

Finally, one of the most difficult problem is the blind source separation. Indeed, the microphone records sounds that are often simultaneous as showed by the high number of mixed sounds in our experiment. Some techniques developed in other areas of signal processing may be considered to

TABLE III
EVERY DAY LIFE SOUND CORPUS

Category	Sound Classe	Sound Nb.	Mean SNR (dB)	Mean length (ms)	Total length (s)
Human sounds:					
	Cough	8	14.6	79	0.6
	Fart	1	13	74	0.01
	Gargling	1	18	304	0.3
	Hand Snapping	1	9	68	0.01
	Mouth	2	10	41	0.01
	Sigh	12	11	69	0.8
	Song	1	5	692	0.7
	Speech	669	11.2	435	290.8
	Throat Roughing	1	6	16	0.02
	Whistle	5	7.2	126	0.6
	Wiping	4	19.5	76	0.3
Object handling:					
	Bag Frisking	2	11.5	86	0.1
	Bed/Sofa	16	10	15	0.2
	Chair Handling	44	10.5	81	3
	Chair	3	9	5	0.01
	Cloth Shaking	5	11	34	0.1
	Creaking	3	8.7	57	0.1
	Dishes Handling	68	8.8	70	4.7
	Door Lock&Shutting	278	16.3	93	25
	Drawer Handling	133	12.6	54	7
	Foot Step	76	9	62	4
	Frisking	2	7.5	79	0.1
	Lock/Latch	162	15.6	80	12.9
	Mattress	2	9	6	0.01
	Object Falling	73	11.5	60	4.4
	Objects shocking	420	9	27.6	11.6
	Paper noise	1	8	26	0.03
	Paper/Table	1	5	15	0.01
	Paper	1	5	31	0.03
	Pillow	1	5	2	0
	Rubbing	2	6	10	0.02
	Rumbling	1	10	120	0.1
	Soft Shock	1	7	5	0
	Velcro	7	6.7	38	0.2
Other:					
	Mixed Sound	164	11	191	31.3
	unknown	231	8.5	25	5.8
Outdoor sounds:					
	Exterior	24	10	32	0.77
	Helicopter	5	10	807	4.4
	Rain	3	6	114	0.3
	Thunder	13	7.5	208	2.7
Device sounds:					
	Bip	2	8	43	0.08
	Phone ringing	69	8	217	15
	TV	1	10	40	0.04
Water sounds:					
	Hand Washing	1	5	212	0.2
	Sink Drain	2	14	106	0.2
	Toilet Flushing	20	12	2833	56.6
	Water Flow	13	7	472	6.1
Overall sounds except speech		1886	11.2	107.8	203.3
Overall speech		669	11.2	435.0	291.0
Overall		2555	11.2	193.5	494.3

analyze speech captured with far-field sensors and develop a Distant Speech Recogniser (DSR): blind source separation, independent component analysis, beam-forming and channel selection. Some of these methods use simultaneous audio signals from several microphones.

2) Audio Categorisation: As stated in the beginning of this paper, two main categories of audio analysis are generally targeted: daily living sounds and speech. These categories represent completely different semantic information and the techniques involved to process of these two kinds of signal are quite distinct. However, the distinction can be seen as artificial. The results of the experiment showed a high confusion between speech and sounds with overlapped spectrum. For instance, one problem is to know whether scream or sigh must be classified as speech or sound. Moreover, mixed sounds can be composed of speech and sounds. Several other orthogonal distinctions can be used such as voiced/unvoiced, long/short, loud/mute etc. These would imply using some other parameters such as sound duration, fundamental frequency and harmonicity. In our case, most of the poor results can be explained by the lack of examples used to learn the models and the fact that no reject class has been considered. But choosing the best discrimination model is still an open question.

3) Everyday Living Sounds Recognition: Everyday living sounds identification is particularly interesting for evaluating the distress situation in which the person might be. For instance, window glass breaking sound is currently used in alarm device. A more useful application for daily living assistance is the recognition of devices functioning such as the washing machine, the toilet flush or the water usage [2] in order to assess how a person copes with her daily house duty. Health status can also be assessed by detecting coughing, respiration and other related signs. Another ambitious application would be to identify human nonverbal communication to assess mood or pain in person with dementia. Classifying everyday living sounds in smart home is a pretty recent trend in audio analysis. Due to its infancy, the "best" features to describe the sounds and the classifier models are far from being standardized [5], [11], [14]. Most of the current approaches are based on probabilistic models acquired from corpus. But, due the high number of possible sounds, acquiring a realistic corpus allowing the correct classification of the emitting source in all conditions inside a smart home is a very hard task. Hierarchical classification based on intrinsic sound characteristics (periodicity, fundamental frequency, impulsive or wide spectrum, short or long, increasing or decreasing) may be a way to improve the processing and the learning. Another way to improve classification and to tackle ambiguity, is to use the other data sources present in the smart home to assess the current context. The intelligent supervision system may then use this information to associate the audio event to an emitting source and to make decisions adapted to the application.

4) Adaptation of the Speech Recognition to the Speaker: Speech recognition is a old research area which has reached some standardization in the design of an ASR. The most direct application is the ability to interact verbally with the smart home environment (through direct vocal command or dialog) providing high-level comfort for physically disabled or frail persons. But speech recognition could also play an

important role for the assessment of person with dementia. Indeed, one of the most tragic symptoms of Alzheimer's disease is the progressive loss of vocabulary and ability to communicate. Constant assessment of the verbal activity of the person may permit to detect important phases in dementia evolution. However, before conducting such experimentations, many challenges must be addressed to apply ASR to ambient assisted living.

The public concerned by the home assisted living is aged, the adaptation of the speech recognition systems to aged people in thus an important and difficult task. Experiments on automatic speech recognition showed a degradation of performances with age and also the necessity to adapt the models used to the targeted population when dealing with elderly people. A recent study had used audio recordings of lawyers at the Supreme Court of the United States over a decade [15]. It showed that the performances of the automatic speech recognizer decrease regularly as a function of the age of the person but also that a specific adaptation to each speaker allow to obtain results close to the performances of the young speakers. However, with such adaptation, the model tends to be too much specific to one speaker. That is why Renouard *et al.* [16] suggest to use the recognized word in on-line adaptation of the models. This proposition was made in the assisted living but seems to have been abandoned. An ASR able to recognize numerous speakers requires to record more than 100 speakers. Each record takes a lot of time because the speaker is quickly tired and only few sentences may be acquired during each session. Another solution is to develop a system with a short corpus of aged speakers (i.e. 10) and to adapt it specifically to the person who will be assisted.

It is important to recall that the speech recognition process must respect the privacy of the speaker. Therefore the language model must be adapted to the application and must not allow the recognition of sentences not needed for the application. An approach based on keywords may thus be respectful of privacy while permitting a number of home automation orders and distress situations being recognized. Regarding distress, an even higher level approach may be to use only information about prosody and context.

VI. Conclusion and Future Work

Audio processing (sound and speech) has great potential for health assessment and assistance in smart home such as improving comfort via vocal command and security via distress situations detection. However, many challenges in this domain need to be tackled to make audio processing deployed in assisted living applications. The paper presents the issues in this domain based on two experiments conducted in a health smart home involving the audio processing software AuditHIS. The first experiment was related to distress detection from speech. Most of the encountered problems were due to noise or environmental perturbation. The second experiment was related to the audio analysis of usual daily activities performed by fifteen healthy volunteers. The dataset was recorded in realistic conditions and underlines the main challenges that audio analysis must address in the context of ambient assisted living. Among the most problematic issues were the uncontrolled recording condition, the mixing of audio events, the high variety of different sounds and the complexity to discriminate them. Regarding the latter, we plan to conduct several studies to determine what the most interesting features for sound classification are as well as how hierarchical modeling can improve the classification. Moreover, regarding speech recognition, probabilistic models need to be adapted to the aging population. We are currently recording seniors' voice to adapt our ASR to this population.

References

[1] M. Chan, D. Estève, C. Escriba, and E. Campo, "A review of smart homes- present state and future challenges.," *Computer Methods and Programs in Biomedicine*, vol. 91, pp. 55–81, Jul 2008.

[2] A. Ibarz, G. Bauer, R. Casas, A. Marco, and P. Lukowicz, "Design and evaluation of a sound based water flow measurement system," in *Smart Sensing and Context*, vol. 5279/2008 of *Lecture Notes in Computer Science*, pp. 41–54, Springer Verlag, 2008.

[3] J. Chen, A. H. Kam, J. Zhang, N. Liu, and L. Shue, "Bathroom activity monitoring based on sound," in *Pervasive Computing* (S. B. . Heidelberg, ed.), vol. 3468/2005 of *Lecture Notes in Computer Science*, pp. 47–61, 2005.

[4] D. Litvak, Y. Zigel, and I. Gannot, "Fall detection of elderly through floor vibrations and sound," in *Proc. 30th Annual Int. Conference of the IEEE-EMBS 2008*, pp. 4632–4635, 20–25 Aug. 2008.

[5] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letter*, vol. 24, no. 15, pp. 2895–2907, 2003.

[6] O. Kumico, M. Mitsuhiro, E. Atsushi, S. Shohei, and T. Reio, "Input support for elderly people using speech recognition," tech. rep., Institute of Electronics, Information and Communication Engineers, 2004.

[7] M. Fezari and M. Bousbia-Salah, "Speech and sensor in guiding an electric wheelchair," *Automatic Control and Computer Sciences*, vol. 41, pp. 39–43, Feb. 2007.

[8] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, "Robust environmental sound recognition for home automation," *IEEE Trans. on Automation Science and Engineering*, vol. 5, pp. 25–31, Jan. 2008.

[9] M. Gorham-Rowan and J. Laures-Gore, "Acoustic-perceptual correlates of voice quality in elderly men and women," *Journal of Communication Disorders*, vol. 39, pp. 171–184, 2006.

[10] M. Vacher, A. Fleury, F. Portet, J.-F. Serignat, and N. Noury, *New Developments in Biomedical Engineering*, ch. Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living, pp. 645 – 673. Intech Book, Feb. 2010. ISBN: 978-953-7619-57-2.

[11] M. Vacher, J. Serignat, S. Chaillol, D. Istrate, and V. Popescu, *Speech and Sound Use in a Remote Monitoring System for Health Care*, vol. Lecture Notes in Artificial Intelligence, 4188/2006, pp. 711–718. Springer Berlin/Heidelberg, 2006.

[12] A. Fleury, M. Vacher, and N. Noury, "SVM-based multimodal classification of activities of daily living in health smart homes: Sensors, algorithms, and first experimental results," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 274–283, 2010.

[13] F. Portet, A. Fleury, M. Vacher, and N. Noury, "Determining useful sensors for automatic recognition of activities of daily living in health smart home," in *Intelligent Data Analysis in Biomedicine and Pharmacology*, (Verona, Italy), pp. 63–64, Jul. 19 2009.

[14] H. D. Tran and H. Li, "Sound event classification based on feature integration, recursive feature elimination and structured classification," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, pp. 177–180, 19–24 Apr. 2009.

[15] R. Vipperla, S. Renals, and J. Frankel, "Longitudinal study of asr performances on ageing voices," in *The 9thAnnual Conference of the International Speech Communication Association, INTERSPEECH'08 Proceedings*, (Brisbane, Australia), pp. 2550–2553, 2008.

[16] S. Renouard, M. Charbit, and G. Chollet, "Vocal interface with a speech memory for dependent people," *Independent Living for Persons with Disabilities*, pp. 15–21, 2003.